S C H E D A E   I N F O R M A T I C A E
VOLUME 17/18 2009

Neural Networks Application to Reduction of Train Caused Distortions in Magnetotelluric Measurement Data

MARZENA BIELECKA[1], TOMASZ. DANEK[1], MAREK WOJDYŁA[2], GRZEGORZ BARAN[3]
[1]Departament of Geoinformatics and Applied Computer Science, AGH University of Science and Technology, Mickiewicza 30, 30-059 Kraków, Poland
[2]PBG Geophysical Exploration Company Ltd., Jagiellońska 76, 03-301 Warszawa, Poland
[3]Institute of Computer Science, Jagiellonian University
Prof. Stanisława Łojasiewicza 6, 30-348 Kraków, Poland
e-mail: *bielecka@agh.edu.pl, danek9@geolog.geol.agh.edu.pl*

Abstract. Artificial intelligence methods for MT data processing are proposed. Distortions having a complex structure created by external artificial sources such as, for example, passing train were investigate. In the first part of this paper the time intervals with such type of distortions were found by using a special neuronal system. Next for time intervals found in the first stage the measure curve fragment is removed and then it is replied by the fragment created by a trained perceptron. The experiment showed that used method are effective.

Keywords: MT data, neural network.

1. Introduction

Magnetotelluric (MT) is a one of the most popular and commonly used passive geophysical method. It is widely used in deep lithosphere research but recently a growing number of examples of using MT method in hydrocarbon prospecting and earthquake prediction are observed. Magnetotelluric applies measuring of fluctuations in natural electric (E) and magnetic (H) fields in orthogonal directions at the surface of the Earth to determining the conductivity structure of the Earth at depths ranging from tens of meters to several hundreds of kilometers.

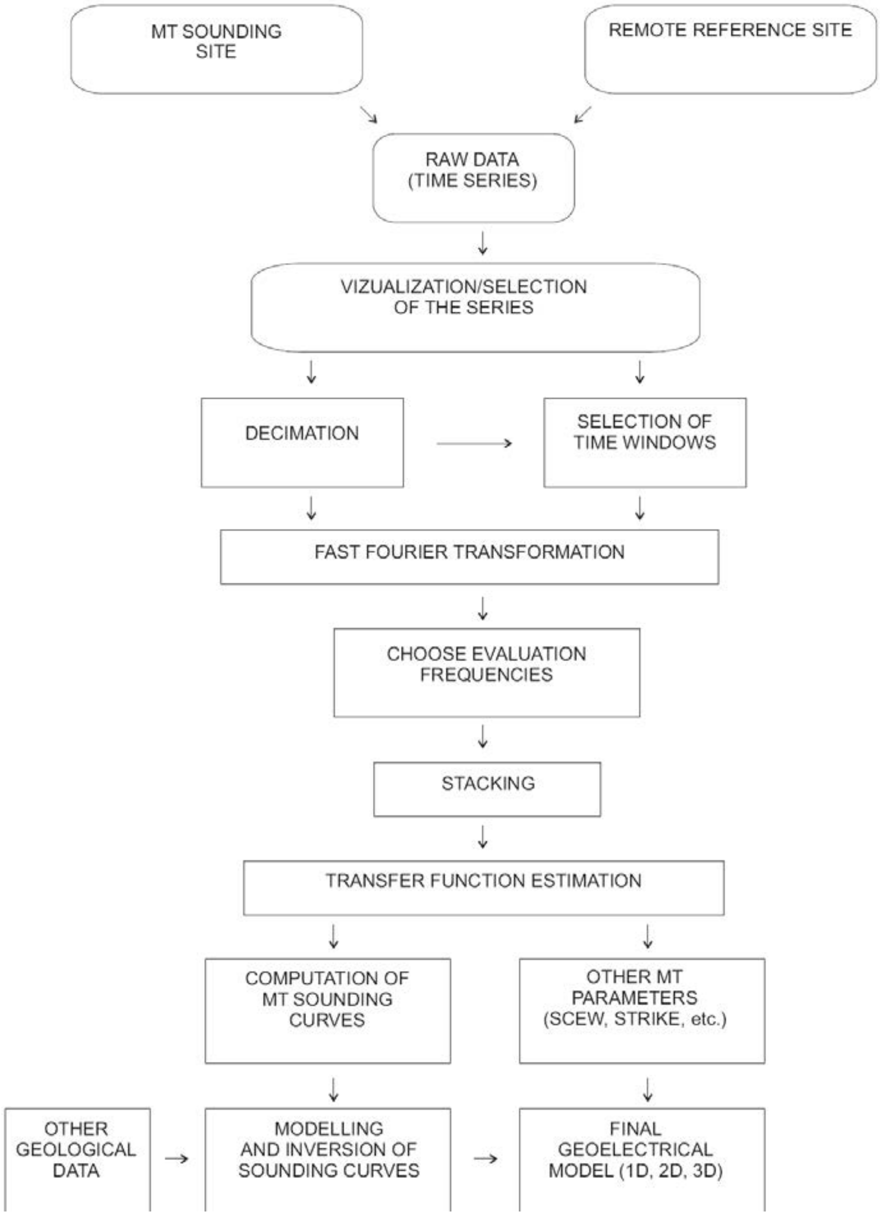

Fig. 1. Simplified scheme of MT processing and interpretation

The fundamental theory of magnetotelluric was first proposed by Tikhonov [1], and Cagniard [2]. Magnetotelluric data, called amplitude and phase sounding curves, are usually obtained by processing of measured time series, which describe electromagnetic variations in particular point on Earth surface. Sounding curves are next interpreted with various inversion methods (Fig. 1). But quality of magnetotelluric data often suffers from high noise level. The most common sources of this noise such as railways, electrical devices, thunderstorms etc. are very hard to avoid especially in urbanized areas. Typically MT signals are recorded in two separated points (measuring point and remote reference point) and then these two records are processed together. Fragments with high noise level are simply removed. But sometimes noise sources like for example DC railways are so intensive that whole record is affected and traditional processing is pointless. In this case, to obtain a solid sounding curve with limited error from raw and very noisy time series other, more effective and sophisticated method of noise reduction is needed. Nowadays the most promising ways of reduction of steady shape intensive noises are based on methods of artificial intelligence [3, 4, 5].

2. Recognition and elimination distortions caused by trains

In order to implement automatic system for train caused distortions recognition and elimination, 121 data sets have been collected. Each set consists of two measurements:

(a) The first one consists of two time series $-E_x(t)$ and $H_y(t)$, without distortions.

(b) The second one consists of two time series $-E_x(t)$ and $H_y(t)$, has been done at the same place, with distortions caused by trains.

The time series consist of 57 till about 5000 measurement points. It was input date for the system.

2.1. Recognition of train caused distortions

The distorted intervals marking in the time series $E_x(t)$ was the first task of the implemented system.

The data analysis showed that distortions caused by trains have a characteristic shape on the time series $E_x(t)$ – see Fig. 2.

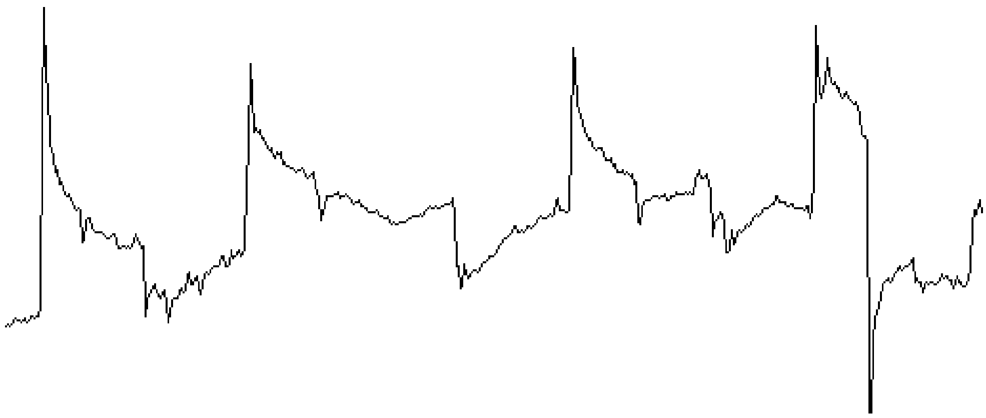

Fig. 2. Train caused distortions on the E_x time series

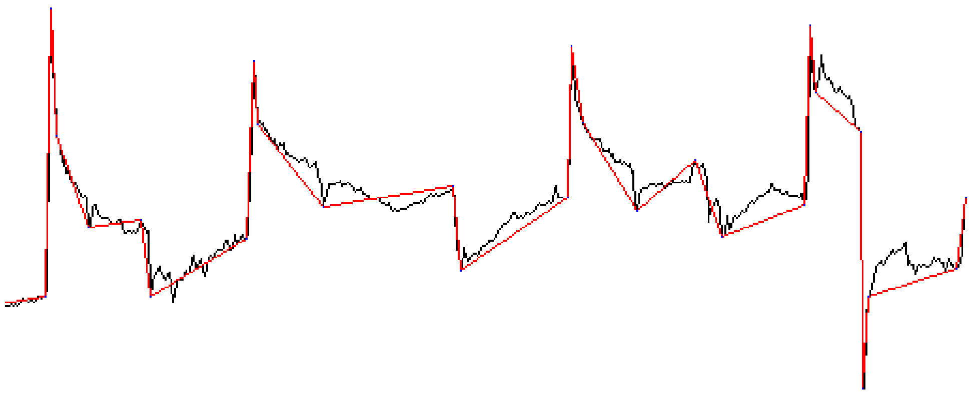

Fig. 3. An example of effect of approximating algorithm using – the time series is approximated by a broken line

2.1.1. Train caused distortions recognition using artificial neural networks (ANNs)

Learning and testing sets were prepared in the following way:

a) The approximating broken line is calculated for each time series $E_x(t)$ (Fig. 3).

b) Sequences of n $(n \geq 1)$ points satisfying a given criterion were chosen.

In this paper three various criteria were used to choose sequences for input data:

$$G_q(p_m,..., p_{m+n}) \Leftrightarrow a(l : p_m \to p_{m+1}) \geq q,$$

$$D_q(p_m,..., p_{m+n}) \Leftrightarrow a(l : p_m \to p_{m+1}) \leq q,$$

$$O_q(p_m,..., p_{m+n}) \Leftrightarrow G_q(p_m,..., p_{m+n}) \vee D_{-q}(p_m,..., p_{m+n}),$$

where: $a(l : p_m \to p_{m+1})$ is a slope of a straight line containing points p_m and p_{m+1}, $d \in R$.

The criteria were applied in order to divide the set S_n into various classes of problems which are solved by specialized ANNs. Thus, the criterion G_q describes sequences in which the first line segment is described by the ascending affined function with great value of differential coefficient whereas D_q corresponds to sequences in which the first line segment is described by the descending affined function with great absolute value of differential coefficient. The criterion O_q describes sequences in which the first line segment is described by the ascending or descending affined function with great absolute value of differential coefficient.

The above criteria were used because train caused distortion begins from big increase or big decrease of the measured value E_x.

c) For each sequence $(p_m,..., p_{m+n}) \in S_n(K)$ input vector was made. It was calculated in two following ways:

$$W_w(p_m,..., p_{m+n}) = [x(p_m \to p_{m+1}), y(p_m \to p_{m+1}),..., x(p_{m+n-1} \to p_{m+n}), y(p_{m+n-1} \to p_{m+n})]$$

$$W_k(p_m,..., p_{m+n}) = [a(l : p_m \to p_{m+1}), d(p_m, p_{m+1}),..., a(l : p_{m+n-1} \to p_{m+n}), d(p_{m+n-1}, p_{m+n})]$$

where:

$x(p_m \to p_{m+1})$ – the first component of the vector $p_m \to p_{m+1}$,

$y(p_m \to p_{m+1})$ – the second component of the vector $p_m \to p_{m+1}$,

$a(l : p_m \to p_{m+1})$ – slope of a straight line to which points p_m and p_{m+1} belong,

$d(p_m, p_{m+1})$ – Euclidean distance between points p_m and p_{m+1}.

d) A number value was assigned to each input vector in dependence on the investigated fragment of the time series $E_x(t)$ was distorted or not:

$f(W) = 1$ if vector W has been obtained from distorted time series $E_x(t)$,

$f(W) = 0$ if vector W has been obtained from undistorted time series $E_x(t)$.

About ¾ elements from the obtained set of pairs $(W, f(W))$ were selected randomly and they constituted the learning set. Other elements were used as testing set.

2.1.2. ANNs architectures

Two types of ANNs were used:

a) Multilayered perceptron with one hidden layer consisting of sigmoid neurons [7]:

$$f(x) = \frac{1}{1+e^{-x}}.$$

The output layer consists of linear neurons.

b) Multilayer network with radial basis neurons in the hidden layer [6]:

$$g(x_1, ..., x_k) = f(r(x_1, ..., x_k)),$$

where:

$r(x) = \beta |x - w|_k^2$,

$x = [x_1, ..., x_k]$ – input vector,

$w = [w_1, ..., w_k]$ – a neuron weights vector,

$|x - w|_k$ – distance in *k*-dimensional Euclidean space,

$\beta > 0$ – a parameter.

An activation function:

$$f(x) = e^{-x}.$$

There are linear neurons in the output layer.

The efficiency of ANNs prediction was measured using *Mean Absolute Error*.

2.1.3. The system recognizing train caused distortions

For a given time series $E_x(t)$ the train caused distortions recognizing algorithm is of the following form:

1. Estimate the time series $E_x(t)$ for a parameter δ given by the user or calculated using statistic methods.
2. For each sequence of n broken line segments:
   a. check if the sequence satisfies one of the established criterion (G_q, D_q or O_q);
   b. check the proper neural network in dependence of which criterion is satisfied; if none is satisfied interrupt the algorithm;
   c. construct input vector according to W_k or W_w;
   d. activate ANN for the obtained input vector;
   e. remember of the output ANN value for each point t_k from the investigated interval.
3. For each point t_k of the graph the output value of the system is equal to arithmetic mean of all ANNs output values obtained in point 2.e.
4. Mark intervals of time series $E_x(t)$ on which the output value of the system is greater than a given bias $0 \le \gamma \le 1$ (default value is equal to 0.5) as intervals with distortions.

The scheme of the system is shown in Fig. 4.

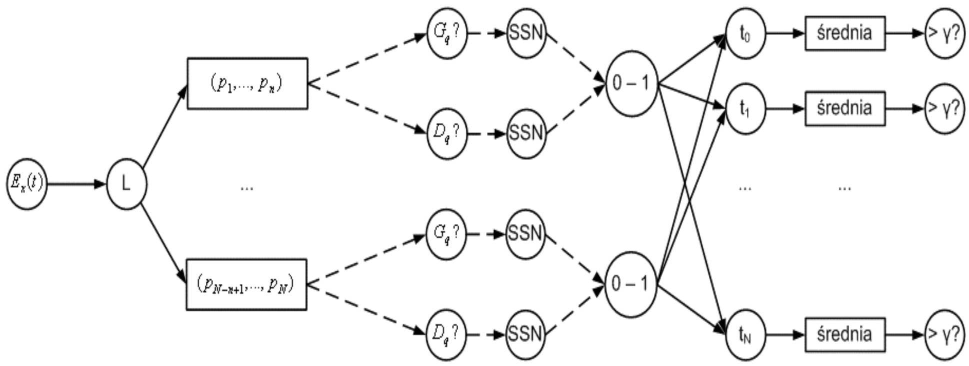

Fig. 4. The scheme of train caused distortions recognizing system using criteria G_q and D_q

2.1.4. Obtained results

The results are shown in Tab. 1. In systems number 4, 8, 9, 11 and 18 two various ANNs were used for line segments sequences satisfying criterion $G_{0,4}$ and $D_{-0,4}$. System 18 was the best one – its efficiency is equal to 75,13%.

2.2. Train caused distortions elimination

The data set consists of 121 subsets. Each of them it is a double measurement of the electric Earth field with and without train caused distortions. The distorted signal was put onto the ANN input. To return a time series like to the corresponding one without distortions was the ANN task.

The ANN has one hidden layer with sigmoidal neurons or radial-based neurons and linear neurons in the output layer. The number of neurons in the output layer was equal to the size of the input vector.

Tab. 1. Distortions recognition results. Notations: RFB – radial based network – neural Network with radial neuron in the hidden layer; WP – multilayer perceptron; δ parameter determined for each time series using statistic methods

The system number	Parameter δ	Input vector size	How the input vector was determined	Criterion	Number of neurons in hidden layer	ANN type	MAE for testing set	Efficiency
1	0,8	20	W_k	$G_{0,4}$	120	RFB	0,3833	72,15%
2	0,8	20	W_k	$G_{0,4}$	6	WP	0,3961	71,81%
3	0,8	30	W_k	$G_{0,4}$	30	WP	0,3621	74,43%
				$D_{-0,4}$	31	RFB	0,3773	
4	0,8	30	W_k	$G_{0,4}$	5	WP	0,3649	73,51%
				$D_{-0,4}$	8	WP	0,3941	
5	1,3	10	W_w	$G_{0,4}$	15	WP	0,368	61,49%
6	1,3	20	W_w	$O_{0,4}$	63	RFB	0,3689	68,01%

The system number	Parameter δ	Input vector size	How the input vector was determined	Criterion	Number of neurons in hidden layer	ANN type	MAE for testing set	Efficiency
7	1,3	20	W_w	$G_{0,4}$	6	WP	0,3912	63,36%
8	1,3	20	W_k	$G_{0,4}$	22	WP	0,3769	68,03%
8	1,3	20	W_k	$D_{-0,4}$	5	WP	0,3877	68,03%
9	1,3	20	W_k	$G_{0,4}$	3	WP	0,39	67,52%
9	1,3	20	W_k	$D_{-0,4}$	5	WP	0,3877	67,52%
10	1,3	30	W_k	$O_{0,4}$	5	WP	0,3727	65,35%
11	1,3	30	W_k	$G_{0,4}$	2	WP	0,3487	66,17%
11	1,3	30	W_k	$D_{-0,4}$	4	WP	0,3669	66,17%
12	1,3	30	W_w	$G_{0,4}$	23	WP	0,3545	67,22%
13	1,3	40	W_w	$O_{0,4}$	18	RFB	0,359	61,59%
14	1,3	40	W_w	$G_{0,4}$	43	WP	0,3396	66,17%
15	1,3	40	W_w	$D_{-0,4}$	102	RFB	0,302	62,03%
16	variable	10	W_k	$G_{0,4}$	6	WP	0,4119	64,79%
17	variable	30	W_k	$O_{0,4}$	46	RFB	0,401	70,46%
18	variable	30	W_w	$G_{0,4}$	50	WP	0,378	75,13%
18	variable	30	W_w	$D_{-0,4}$	8	WP	0,3831	75,13%
19	variable	40	W_w	$G_{0,4}$	51	RFB	0,3901	71,55%
20	variable	40	W_k	$D_{-0,4}$	24	RFB	0,3922	68,82%
21	variable	40	W_k	$D_{-0,4}$	6	RFB	0,3992	68,40%

Learning and testing sets were constructed in the following way:

1. Each time series $E_x(t)$ was divided into fragments consisted of M points.
2. Each obtained fragment $F_k(t)$ was divided into m sub-fragments and the integral of each fragment has been calculated.

$$G_k(t) = \sum_{i=0}^{m-1} F_k(mt+i), \qquad m \geq 1,\ t = 0,1,2,\ldots,\left\lfloor \frac{M}{m} \right\rfloor - 1.$$

3. The data set was composed of input vectors with $\left[G_k(0),\ldots,G_k\left(\left\lfloor \frac{M}{m} \right\rfloor - 1\right)\right]$ and corresponding output vectors $\left[G_k'(0),\ldots,G_k'\left(\left\lfloor \frac{M}{m} \right\rfloor - 1\right)\right]$ without distortions.
4. The ¾ of the data set was used as the learning set and ¼ as the testing set.

The used data it is about 15% of needed data being necessary for learning the ANNs with the highest efficiency. Results are shown in Tab. 2 and Tab. 3. In Tab. 2 there are results that were obtained for the smaller data set consisted of 74 measurement subsets. In Tab. 3 there are result for bigger data set consisted of 121 measurement subsets. It is noticed that for the bigger data set the result is changed significantly.

Tab. 2. Neural networks training results. The networks were used for elimination of train caused distortions. Notations: RFB – radial based network – neural Network with radial neuron in the hidden layer; WP – multilayer perceptron

Index	M	m	Number of neurons in the last layer	ANN type	Number of hidden neurons	MAE for learning set	MAE for testing set	Cardinality of the learning set	Cardinality of the testing set	Weights number in the neural network
1	600	12	50	RFB	12	11,409	19,899	49	16	1200
2	300	12	25	RBF	10	19,89	23,565	142	47	500
3	300	6	50	RFB	11	10,108247	13,112735	142	47	1100
4	150	6	25	WP	20	152,93	179,178	324	108	1000
5	150	6	25	RFB	10	11,003	13,101	324	108	500
6	150	6	25	RFB	21	11,061	12,044	324	108	1050

Tab. 3. Results for ANNs used for train caused distortions elimination. Notations: RFB – radial based network – neural Network with radial neuron in the hidden layer; WP – multilayer perceptron

Index	M	m	Number of neurons in the last layer	ANN type	Number of hidden neurons	MAE for learning set	MAE for testing set	Cardinality of the learning set	Cardinality of the testing set	Weights number in the neural network
1	600	12	50	RFB	6	16,367	24,855	38	12	600
2	300	12	25	RBF	10	22,959	28,115	105	35	500
3	300	6	50	RFB	16	10,020	15,669	105	35	1600
4	150	6	25	RFB	27	11,070	15,294	239	80	1350

3. Concluding remarks

In the case of train caused distortions the results should be regarded as preliminary because the number of learning examples is about 15% of the needed number. However, even is such situation the system achieved 75% efficiency, which, in this context, is a quite good result. In the case of ANNs which removes train caused distortions the results are very promising. To enlarge the learning set of about 50 examples improved outcomes significantly. This probably means that the obtained results can be improved by adding next examples to the learning set.

Acknowledgements. This article was written within the framework of a research grant founded by the Polish Ministry of Science and Higher Education, no N525 029 32 /2768.

4. References

[1] Cagniard L.; *Basic theory of the magnetotelluric method of geophysical prospecting*, Geophysics, 18, 1953, pp. 605–637.

[2] Tikhonov A.N.; *The determination of the electrical properties of deep layers of the Earth's crust*, Dokl. Acad. Nauk. SSR, 73, 1950, pp. 295–297.

[3] Manoj C., Nagarajan N.; *The application of artificial neural networks to magnetotelluric time-series analysis*, Geophysical Journal International, 153, 2003, pp. 409–423.

[4] Moorkamp M., Jones A.G., Rao C.K.; *Processing magnetotelluric time series with adaptive filters and neural networks*, 18th EM Induction Workshop, El Vendrell, Spain, 2006.

[5] Popova I.V., Ogawa Y.; *Application of a modified Hopfield neural network to noisy magnetotelluric data*, Izvestiya Physics of the Solid Earth, Vol. 43,3, 2007, pp. 217–224.

[6] Yee P.V., Haykin S.; *Regularized Radial Basis Function Networks*, Theory and Applications, 2001.

[7] Tadeusiewicz R.; *Sieci neuronowe*, Akademicka Oficyna Wydawnicza, 1993.

Received December 29, 2008